Pré-GDR : axe 7, Exploration de masses de données langagières

Animateurs : Thierry Charnois,Jean-Pierre Chevallet
Contributeurs : voir liste ci-dessous

Version du 19 Juin 2018

1 Introduction

Le but de ce document est de préparer le pré-GDR TAL et de participer à la rédaction de la première version du document préparatoire pour Juin 2018. Cet axe concerne l'exploration et l'exploitation de masses de données langagières. Les travaux préparatoires ont produit une liste de mots clés qui restera à affiner. Cette liste conduit à la structure actuelle de ce document.

- Exploration de masses de données langagières (en interaction avec Madics);
- Recherche d'information, Indexation;
- Systèmes de Recommandation;
- Navigation dans des informations ou des documents;‘
- Synthèse d'informations et de documents;
- Agregation et Extraction d'informations;
- Systèmes Question / Réponse;
- Analyse des réseaux sociaux;
- Fouille d'opinion;
- Profilage de textes;
- Veille documentaire;
- Lexicométrie;

Nous proposons une première esquisse d'une structuration des thématiques de l'axe selon deux directions : les techniques, les outils et les méthodes d'explorations de données textuelles, d'une part, et les applications, domaines et usages. Ces deux directions ne sont pas forcément orthogonales et non définitives. Techniques, outils, méthodes :

- Indexation, Lexicométrie : concerne la pondération, et la représentation de l'information;
- Annotation de documents : au sens de l'association d'une partie de texte à une entrée d'une ressource, comme une ontologie;
- Système de Recherche d'Information;
- Classification, visualisation, présentation, navigation dans des documents;
- Systèmes de Question / Réponse (QA);
- Extraction d'information;
- Systèmes de recommandation;
- Analyse des réseaux sociaux;

Applications, domaines, usages :

- Détection de plagiat
- Recherche d'information, recherche de réponses à des questions précises;
- Veille;
- Synthèse et agrégation de documents;
- Cartographie d'un domaine : usage des termes, des concepts, évolution temporelle, tendances, visualisation;
- Recommandation;
- Profilage de textes, attribution d'auteurs, personnalisation linguistique;
- Fouille d'opinion (sentiments, etc ...).

2 Participants

Cette partie liste les participants actuels qui ont eu le temps de participer aux discussions et à la rédaction de la version actuelle de ce document. Elle doit être complétée pour mieux couvrir tous les thèmes et mots clés présentés dans la partie précédente.

Patrice Bellot <patrice.bellot@lis-lab.fr>,
Catherine Berrut <Catherine.Berrut@imag.fr>,
Romaric Besançon <romaric.besancon@cea.fr>,
Davide Buscaldi <davide.buscaldi@lipn.univ-paris13.fr>,
Peggy Cellier <peggy.cellier@irisa.fr>,
Charlet Jean <jean.charlet@upmc.fr>,
Max Chevalier <Max.Chevalier@irit.fr>,

Vincent Claveau <vincent.claveau@irisa.fr>,
Gaël Dias <gael.dias@unicaen.fr>,
Stéphane Ferrari <stephane.ferrari@unicaen.fr>,
Brigitte Grau <bg@limsi.fr>,
Lynda Tamine Lechani <Lynda.Lechani@irit.fr>,
Anne-Laure Ligozat <annlor@limsi.fr>,
Fabrice Maurel <fabrice.maurel@unicaen.fr>,
Josiane Mothe <josiane.mothe@irit.fr>,
Patrick Paroubek <pap@limsi.fr>,
Solen Quiniou <solen.quiniou@univ-nantes.fr>,
Mathieu Roche <mathieu.roche@cirad.fr>,
Didier Schwab <didier.schwab@univ-grenoble-alpes.fr>,
Lynda Tamine Lechani <Lynda.Lechani@irit.fr>

3 Contributions de la communauté

Chaque thématique identifiée débute par la description de la problématique et par état des lieux. Elle contient ensuite un résumé des grands enjeux. Elle debrait également proposer un positionnement par rapport à d'autres structures (autre GDR, autre insitut...), ainsi que des fait marquant en 2018 (médiation scientifique, réalisation grand public, lien avec l'industrie).

Cette version est préliminaire et encore incomplète au niveau de la couverture des thèmes abordés.

3.1 Système de Recherche d'Information

Participants : Catherine Berrut, Lynda Tamine Lechani, Jean-Pierre Chevallet, Mathieu Roche, Gaël Dias

3.1.1 Problématique et état des lieux

Les Systèmes de Recherche d'Information (SRI) ont pour objectif le filtrage et la sélection des documents qui contiennent ce que l'utilisateur pourra extraire comme information répondant à son besoin. Les SRI sont donc des médiateurs numériques pour faciliter l'accès à des masses de documents. La Recherche d'Information dans des documents textuels est donc par définition 'sémantique'. Les systèmes actuels permettent cet accès à l'information via des modèles statistiques, représentant, telle une signature numérique, la sémantique du contenu des documents. Dans la suite, nous nous intéressons exclusivement aux SRI textuels. Les questions que nous posons alors sont les suivantes :

- Quelle "sémantique" est utile et exploitée par les SRI ?
- Quelles sont les ressources textuelles qui sont utiles et effectivement utilisées dans les SRI (ex: ontologie, terminologie, thésaurus) ?

- Quel est l'utilisation effective des ressources dans les SRI ?
- Comment évaluer l'impact de l'usage de ces ressources sur la qualité finale du SRI en terme de satisfaction de l'utilisateur ?

La notion de "sémantique" dans le contexte des SRI est difficile à cerner. Néanmoins, on peut la cerner par le besoin de combler le fossé sémantique qui existe entre la requête de l'utilisateur et le document lié à la *disparité du vocabulaire*. Le phénomène est identifié par la notion de 'term mismatch'.

Bast et al. (2016) présentent la RI comme le point d'intersection entre les documents textuels et la recherche de mots-clés, alors que, par exemple, les systèmes de question-réponse se positionne comme l'intersection entre les documents textuels et l'interrogation en langue naturelle. La notion même de sémantique en RI est difficile à définir, d'autant que les modèles ne l'explicitent pas et l'approchent principalement que par des statistiques : les récents modèles à base de "word embeddings" (plongement de mots) en sont un exemple phare.

Ressources : Plusieurs types de ressources sont identifiables. Le thésaurus est une ressource linguistique à priori dédié aux SRI, mais en pratique toutes les ressources structurant des termes ont été utilisées dans les SRI.

On note sur ce sujet une forte évolution du domaine notamment pour la construction de ressources. Dans les années 80, les thésaurus jusqu'ici construits manuellement ont été construits automatiquement, par des algorithmes à base de statistiques, de co-occurences, de graphes. Ces premiers algorithmes utilisaient l'hypothèse reprise actuellement qui suppose que les contextes de 2 mots similaires sont similaires. On peut noter également les ressources calculées à partir des textes comme les thésaurus d'associations, basé sur la cooccurrence des mots dans les textes. D'autres approches, dédiées généralement à des domaines pour experts, embarquaient la connaissance dans le langage d'indexation : les applications - essentiellement médicales - évoluant en mode fermé.

Plus récemment les plongements de mots, basés sur le même principe, construisent un vecteur utilisé ensuite pour calculer des distances sémantiques.

Il existe actuellement quatre grandes familles de ressources :

- les ressources de domaine (en médical UMLS, Gene Ontology, etc),
- les ressources générales (wikipedia, dbpedia, garo, geonames, wornet, etc),
- les ressources issues de service web (foursquare, telp, etc),
- les ressources pour la sémantique distributionnelle.

Ces familles de ressources peuvent se conjuguer.

Modèles : La notion de *modèle de RI* fixe le cadre de la définition d'un SRI : il s'agit de décrire la représentation interne des requêtes et des documents, ainsi que la manière qu'a le système d'estimer la pertinence d'un document par rapport à cette requête. Le modèle le plus répandu est celui du *sac de mots*, qui

conduit aux modèles vectoriels ou probabilistes (ex: modèle de langue). Peu d'études conduisent à des directions différentes, comme les modèles à base de formules logiques, comme les logiques de descriptions ou les graphes conceptuels. Une des difficultés des modèles à base de logique est la place des probabilités, ou plus généralement de l'incertitude. L'objectif d'un SRI étant un classement des documents par ordre de pertinence, il faut nécessairement des critères non binaires pour réaliser ces classements.

La question posée est également la place des ressources dans ces modèles. Elles se cantonnent à la modification des requêtes (pour limiter le 'term mismatch'), parfois des documents. Les ressources sont très rarement mise à contribution dans la fonction de correspondance. On distingue 3 grandes approches : les approches orientées appariement, les approches orientées ordonnancement, les approches orientées apprentissage.

Une direction particulière consiste à l'utilisation d'un réseau de neurones profond (deep learning), comme étape supplémentaire d'un SRI plus classique pour améliorer le classement d'un sous ensemble des documents de la collection. Cette approche n'est possible que si l'on dispose d'un nombre important de requêtes résolues (couple requête, document pertinent).

Exploitation des ressources : En première approche, la RI sémantique est celle qui exploite des ressources linguistiques. Il y a trois approches différentes dans l'exploitation de ces ressources :

- Les approches orientés appariement : l'objectif est d'améliorer la représentation des requêtes ou des documents pour augmenter les chances d'appariement. Cela peut se faire par :
  - les approches à base d'expansion de la requête : on reformule la requête à l'aide de mots / concepts pertinents, ou on pondère les termes de la requête;
  - les approches à base d'expansion de document : on procède de la même manière, mais cette expansion peut être statique, i.e. indépendante de la requête, ou dynamique, en tenant compte de la requête ou du retour utilisateur;
  - les approches à base d'expansion de la requête et des documents.
- les approches orientées ordonnancement;
- les approches orientées apprentissage : il s'agit d'apprendre la fonction de pertinence, ou d'apprendre un représentation des mots / concepts / documents. Cela conduit à deux sous domaines : le 'learning to rank' et les approches de correspondance par des réseaux de neurones.

3.1.2 Verrous

Au niveau des verrous actuels, on constate l'absence d'un guide de bonnes pratiques pour assurer la robustesse de tâches en RI sémantique. On constate

également la difficulté de délimiter la frontière entre la qualité des ressources et qualité de leur usage en RI sémantique. Enfin la question se posent des niveaux de sémantiques élémentaires nécessaires. Pour le futur, on peut s'interroger sur les verrous suivants :

- Vers des niveaux de semantique plus elevé : il faut aller vers une semantique du dialogue (systemes conversationnels), semantique de la tache/session (systemes orientes taches/decision, systemes de RI aggregative generant des objets a valeur ajoutee GIO etc)
- Vers une semantique transparente et explicable : Au dela de l'utilisation de representations textuelles "semantisees", s'interroger sur la qualite des "inferences" semantiques qui ont permis d'aboutir aux resultats pour les expliquer a l'utilisateur (Fairness and transparency in IR)
- Besoin de croiser les regards 'scientifiques' sur la question de la semantique dans différents domaines de l'informatique : RI, Machine Learning, Gestion de Connaissances, NLP : quelle semantique ? aboutit-on a des 'sens' alignables ? Quelles differences et pourquoi ?

3.2 Systèmes de Question / Réponse (QA)

Participants : Brigitte Grau, Anne Laure,

3.2.1 État des lieux

La recherche de réponse à des questions vise à retourner une réponse précise à une question posée en langage naturel. Cette réponse peut être extraite d'un corpus de documents textuels, ou d'une base de connaissances, en domaine ouvert ou dans un domaine de spécialité. Les problématiques sont d'une part de rapprocher deux formulations de la même information (paraphrases si textuel, langage naturel vers représentation sémantique si structuré), et d'autre part, d'extraire la réponse. Côté bases de connaissances, les systèmes de question-réponse effectuent généralement une analyse sémantique par apprentissage supervisé (Yu et al., 2017). Côté textuel, les systèmes comportent généralement une phase de recherche d'information, qui va chercher à sélectionner les passages les plus pertinents par rapport à la question, et une phase d'extraction d'information, où la réponse précise va être extraite de ces phrases. Cette dernière étape peut comporter la détection et désambiguïsation des entités et relations de la questions.

Les travaux actuels se focalisent notamment sur l'ordonnancement des passages réponse (Chen and Van Durme, 2017) puis le rapprochement entre question et passage et l'extraction de la réponse (Wang et al., 2017). Certains systèmes combinent les informations structurées et textuelles afin de répondre à la question (Savenkov and Agichtein, 2017; Das et al., 2017). La majorité des méthodes sont fondées sur un apprentissage supervisé par des réseaux de

neurones profonds, pour aligner des textes ou pour apprendre des relations et entités dans des graphes.

Pour un état de l'art plus complet, voir (Grau et al., 2015).

La tâche de QA évolue également vers du machine reading: étant donné un texte, répondre à des questions sur ce texte. La problématique n'est plus axée sur la recherche des textes pertinents, mais plutôt sur les aspects compréhension d'un texte par la machine et sa capacité à abstraire et faire des inférences pour répondre.

3.2.2 Enjeux

- aspects hybrides: avec le développement des données liées sur le web, en domaine général ou en domaine de spécialité, se pose le problème d'intégrer texte et connaissances structurées.
- question/réponse visuel: combiner informations textuelles ou structurées avec informations contenues dans des images.
- questions longues/complexes: les jeux de questions actuels comportent beaucoup de questions se traduisant par une seule relation de la base de connaissance (à une date ou un lieu près)
- constitution de ressources de taille suffisante pour appliquer des méthodes par réseaux de neurones et de qualité.
- QA interactif: intégration QA et système de dialogue avec prise en compte du contexte dialogique.

3.3 Extraction d'information

Participants : Mathieu Roche, Thierry Charnoit, Peggy Cellier, Anne-Laure Ligozat, Romaric Besançon

3.3.1 Problématique

L'extraction d'information (EI) est une tâche relativement ancienne qui vise non pas à appréhender un texte dans son entier mais à en extraire des éléments "pertinents". Elle est généralement structurée en plusieurs sous-tâches selon la nature de l'information à extraire: la reconnaissance d'entités nommées, l'extraction de relations binaires entre entités, l'extraction d'événements, qui associent plusieurs entités à des rôles spécifiques dans un événement et l'extraction de relations entre événements, par exemple temporelles ou causales. L'extraction d'information est utilisée aussi bien comme processus de création ou d'enrichissement de bases de connaissances et de ressources linguistiques que comme préalable à d'autres tâches de TAL ou de recherche d'information.

Cette tâche couvre deux cas d'usage, selon que le type des informations à extraire est prédéfini ou non. Dans le premier cas, plus fréquent, on cherche à

repérer dans des documents des instances d'entités, de relations ou d'événements connus. Le second cas correspond plus à une tâche d'exploration non guidée des données textuelles, dans laquelle le type même des entités, relations ou événements est découvert de façon automatique à partir des documents (e.g. Chambers, 2011).

Plusieurs types d'approches s'intéressent au problème de l'extraction d'information. Ces approches peuvent être découpées en 3 grandes familles : les méthodes statistiques, les méthodes symboliques et les méthodes hybrides. Le développement de ressources annotées a permis l'essor d'approches à base d'apprentissage supervisé comme lors des campagnes MUC et ACE (Hobbs et Riloff, 2010), et plus récemment les campagnes TAC.

3.3.2 Enjeux

Précision des informations extraites Si la tâche de reconnaissance des entités nommées pour les types généraux est arrivée à un stade de maturité, les enjeux actuels pour les entités nommées portent sur l'augmentation de la précision des informations extraites: types d'entités plus fins ("chanteur", "acteur", "homme politique" au lieu de "personne"), nature compositionnelle des entités (par exemple "nom", "prénom", "titre" pour une personne), désambiguïsation des entités nommées (*Entity Linking*).

Par ailleurs, le développement de méthodes d'extraction de relations ou d'événements reste encore un verrou important (par exemple, l'état de l'art pour la reconnaissance des mentions d'événements dans la campagne TAC KBP 2017 est à 56% de F-score).

Degré de supervision selon la disponibilité des données annotées Un des défis de l'extraction d'information est la quantité de données annotées disponibles, qui se traduit en pratique par la mise en place de méthodes avec différents niveaux de supervision:

- des méthodes complètement supervisées, utilisant en général des modèles neuronaux dans les approches les plus récentes;
- des méthodes faiblement supervisées, comme les modèles de supervision distante, dans lesquelles des données textuelles annotées sont créées automatiquement en exploitant des bases de connaissances;
- des méthodes non supervisées, comme la problématique de l'*Open Information Extraction* (*OpenIE*) (Banko and al., 2007; Del Corro and Gemulla, 2013), qui s'affranchit de la nécessité de disposer de corpus étiquetés et vise l'indépendance au domaine: le type des relations à extraire n'est, dans ce cas, pas limité à un ensemble pré-défini, ce qui permet de traiter la diversité des relations existantes en domaine ouvert.

Adaptation au domaine ou au genre Un autre défi, lié à la fois à la disponibilité de données annotées et à la profusion des données textuelles disponibles,

est celui de l'adaptation automatique des modèles d'extraction d'information au domaine ou au genre, pour prendre en compte l'hétérogénéité des textes qui se manifeste à divers niveaux : des genres textuels, des registres de langue dont l'oralité, des réseaux sociaux avec les dialectes, les sociolectes, etc.

3.4 Systèmes de recommandation

Participants : Max Chevalier, Patrice Bellot

3.4.1 Problématique

Les systèmes de recommandation visent à proposer automatiquement à l'utilisateur des items (musique, film, produit, information...) d'intérêt sans qu'il ait nécessairement à interroger le système. Le fonctionnement d'un tel système repose sur une approche essentiellement "push", c'est-à-dire que les items sont "poussés" vers l'utilisateur automatiquement. Le fonctionnement des systèmes de recommandation repose sur des approches, techniques et outils partagés pour un grand nombre avec la recherche d'information.

Pour atteindre cet objectif, les systèmes de recommandation reposent le plus souvent sur une analyse des activités (e.g. des traces) ainsi que des retours explicites qui lui sont fait (retour de pertinence par exemple via une note, une évaluation textuelle..).

En parallèle d'autres types de systèmes de recommandation, que l'on pourrait appeler systèmes de suggestion, sont guidés par des requêtes dans lesquelles les caractéristiques des items qui seraient appréciés et de ceux qui ne sont pas désirés. Ces requêtes peuvent être exprimées via des formulaires mais aussi en langue naturelle. Lien fort entre la question de la recommandation et celle de l'analyse de sentiments.

3.4.2 Enjeux

Alors que le TALN a été intégré depuis plusieurs années dans le monde des systèmes de recommandation, les enjeux actuels tournent autour de deux dimensions spécifiques :
- Le TALN appliqué à la voix. En effet, l'essor de toutes les technologies "voice control" telles que les assistants vocaux nomades et personnels proposent une nouvelle dynamique pour les systèmes de recommandation
- Même si c'est une thématique qui possède déjà un petit historique, un des enjeux consiste toujours à identifier, via le TALN, le sentiment d'un utilisateur vis-à-vis d'un item (Sentiment Analysis ou encore Opinion Mining). Cet élément est très important car il permet l'interprétation fine d'un contenu textuel produit par l'usager (User Generated Content). Ces contenus peuvent être des avis, des retours de tests etc... Il est à noter que ces deux enjeux peuvent être couplés..

Un autre enjeu concerne l'identification des caractéristiques des items à partir de leur contenu ou des commentaires qui les décrivent sur les réseaux sociaux.

Par rapport à la RI non orientée recommandation la principale caractéristique est la prise en compte d'avis (des notes ou pour le lien avec le TAL des commentaires écrits) Cela implique :

- l'analyse de sentiments sur les commentaires des utilisateurs et un sous-problème : comment trouver des commentaires (des critiques) pertinentes ?
- l'analyse de sentiments et des émotions sur les contenus eux-mêmes (ex. s'agit-il d'un livre joyeux ou triste ?)
- l'analyse des aspects associés aux sentiments (quelles sont les caractéristiques négatives / positives) = Aspect Based Sentiment Analysis (ABSA)
- la mise en relation des caractéristiques souhaitées / des caractéristiques trouvées (recherche d'information et facettes, pertinence multi-dimensionnelle)
- la mise en relation des items (Web des items)
- la complémentarité des contenus et des commentaires associés (cas typique : CLEF Social Book Search)

De façon générale sont alors considérés :

- le traitement de requêtes longues en langue naturelle
- les liens entre analyse de sentiments et d'émotion
- la pertinence multi-dimensionnelle et notamment : niveau linguistique des contenus vs des lecteurs etc.
- l'adaptation au contexte (recommandation pour groupe vs individu isolé, âge de l'utilisateur, handicaps langagiers etc.)

D'autres types de recommandations sont en lien étroit avec le TAL : celui de la suggestion d'éléments linguistiques ou assimilés (par ex. émojis). La question est : quel mot proposer à la suite de mots déjà saisis par l'utilisateur ? (suggestion et recommandation de requêtes, complétion automatique...).

3.5 Analyse des réseaux sociaux

Participants : Josiane Mothe, Davide Buscaldi, Patrice Bellot, Gaël Dias, Romaric Besançon

3.5.1 Problématique

L'analyse des réseaux sociaux porte sur (1) les contenus échangés (2) les liens entre individus ou communautés. Les applications sont variées. Il peut s'agir de détecter un changement chez des individus (détection de la depression dans le challenge e-risk, détection de l'agressivité dans le challenge TRAC, ...), d'extraire des communautés échangeant sur les mêmes sujets, de prédire la diffusion d'information dans le réseau étudié, d'étudier les influences sur un sujet.

3.5.2 Enjeux

Les enjeux de cette problématique sont liés à cette double information, à la fois sur le contenu textuel échangé et sur la structure des interactions sociales entre les individus:

- exploiter les informations textuelles pour améliorer l'analyse des interactions sociales: utiliser des représentations sémantiques des contenus textuels dans les algorithmes d'analyse des graphes d'interactions;
- exploiter, de façon duale, les informations sociales pour améliorer l'analyse du contenu textuel: par exemple, utiliser la détection de communautés pour connaître le point de vue d'une personne sur un sujet et mieux analyser son discours (détection de l'ironie, prise en compte de la relativité du sens d'un mot selon le point de vue);
- analyser le contenu textuel des réseaux sociaux en traitant les spécificités de cette langue: les modèles de TAL développés pour la langue générale sont souvent bien moins performants sur la langue des réseaux sociaux et doivent être adaptés (une possibilité est, par exemple, de considérer la langue des tweets comme une langue étrangère ou un dialecte proche de la langue originale et d'utiliser des techniques de projection interlingue pour améliorer les modèles). Cette analyse doit aussi prendre en compte des éléments spécifiques à cette langue, comme les hashtags, les emojis etc.
- développer des méthodes qui permettent de gérer l'hétérogénéité des données: la nature des interactions sociales ou le genre textuel des contenus peuvent être différents selon la nature du réseau social considéré. Pour traiter ce problème, on doit soit construire des méthodes robustes génériques, soit développer des approches pour l'adaptation automatique des modèles à de nouvelles données;
- exploiter les méthodes de fouille de graphes ou de visualisation de données structurées à large échelle, utilisées dans l'analyse des réseaux sociaux, à d'autres types d'informations linguistiques et sémantiques structurées en graphes (par exemple, des réseaux sémantiques de mots)
- graphes multicouches (utilisateurs-auteurs, contenus-items, éditeurs-collections, thèmes-mots-clés-catégories).

3.6 Détection de plagiat

Participants : Didier Schwab

3.6.1 Problématique

Le plagiat est l'appropriation de contenu sans le consentement de son auteur et/ou sans citer ses sources, et ainsi de le présenter comme sa propre œuvre ou création (adapté de (Ferrero, 2017)). Il s'agit d'un problème qui est devenu

crucial en particulier dans le domaine éducatif depuis l'avènement d'Internet et la mise à disposition massive de documents. Plusieurs cas ont fait l'objet d'affaires médiatiques relativement importantes. On peut citer le philosophe des sciences et physicien, Étienne Klein qui dans son livre *Le pays qu'habitait Albert Einstein*, plagierait des textes de plusieurs auteurs dont Zola et Aragon. De même Marine Le Pen entre les deux tours de la présidentielle française 2017 emprunte de nombreuses phrases d'un discours de François Fillon, candidat éliminé au premier tour, deux semaines auparavant.

Plusieurs étude ont montré que le plagiat était un grave problème dans le milieu académique. Guibert and Michaut (2011) montrent que sur une étude menée sur 1485 étudiants français, 34,5% reconnaissaient avoir *"recopié un texte ou une partie d'un texte pour le présenter comme un travail personnel"*. On retrouve des chiffres globalement similaires en Europe (Gibney, 2006) et Amérique du Nord (Josephson Institute, 2011) (McCabe, 2010).

Le plagiat est moins présent dans les publications scientifique, cependant la réutilisation par les coauteurs de portions de leurs propres écrits est un phénomène assez fréquent (Francopoulo et al. (2016)) et l'emploi de procédés de détection automatique se généralise de plus en plus sous la pression des éditeurs et du nombre croissant de canaux de diffusion et de publications.

En ce qui concerne le texte, on considère qu'il y a plusieurs types de plagiat monolingue qui ne sont pas forcément exclusifs :

- La copie exacte lorsque le texte est simplement copié mot à mot ; le plagiat avec modification qui inclut
  - La paraphrase lorsque les mots sont changés mais que la structure globale de la phrase est globalement conservée ;
  - La reformulation lorsque seulement les idées sont conservées ;
  - Le passage d'une langue à une autre autrement appelé plagiat translingue.

La figure 1 présente les différents types de plagiat et certaines méthodes de la littérature pour les détecter.

3.6.2 Verrous

On peut identifier plusieurs verrous importants. Le premier concerne les parties susceptibles d'être plagiées (parties 2 sur la figure 1). En effet, pour des raisons calculatoires, il n'est pas possible de tester toutes les parties d'un document un peu long. Il est ainsi nécessaire d'analyser le texte pour y déceler des indices permettant de savoir s'il est possible qu'il ait plusieurs auteurs et si oui, quelles en sont les limites les plus exactes possibles. L'analyse du style d'écriture, ses évolutions, ceux dans le déroulement des idées dans une analyse non-supervisée pourront être des indices forts de changement d'auteurs.

Le second concerne la recherche d'information translingue (partie 3 sur la figure 1). En effet, il faut trouver tout un ensemble de documents candidats

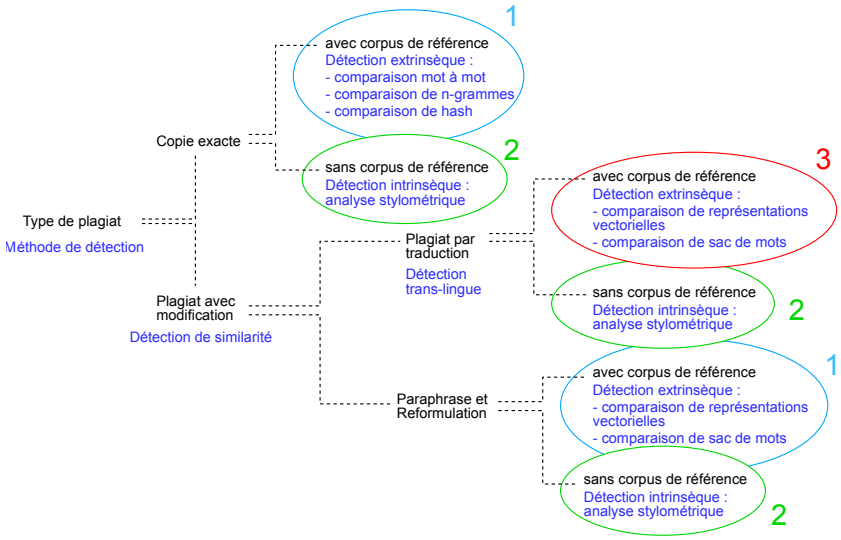

Figure 1: Issu de (Ferrero, 2017), lui-même adapté de la taxonomie de (**?**) des différents types de plagiat et de leurs moyens de détection.

(document éventuellement plagiés) issus de différentes langues. Une piste pourrait être l'utilisation de représentations distribuées distributionnelles continues très multilingues (avec beaucoup de langues) dans un seul espace vectoriel.

Un troisième est l'évaluation de la tâche qui est, comme souvent en traitement des langues, un point crucial. C'est particulièrement vrai en ce qui concerne le plagiat translingue. Des avancées ont été réalisées ces dernières années en particulier avec le corpus de Ferrero et al. (2016) qui permet l'analyse du plagiat translingue à plusieurs granularités (document, phrase, chunck). En revanche, celui-ci ne comprend jamais que 3 langues (français, anglais, espagnol) et il pourrait également être enrichi d'information sur la nature des obfuscations introduites afin de mieux comprendre le comportement des différentes méthodes de détection.

3.7 Fouille d'opinion (sentiments, etc .)

Participants : Patrice Bellot, Davide Buscaldi, Gaël Dias

La fouille d'opinion a fait l'objet d'un volume de travail important au cours de la dernière décennie. Il s'agit d'un sujet qui continue à se développer rapidement dans des nouvelles directions (Liu and Zhang (2012)) et qui trouve des applications dans un grand nombre de domaines. La grande majorité de ce travail a visé le développement de classificateurs de sentiments plus précis, en s'appuyant généralement sur des méthodes supervisées, et plus récemment

en particulier sur des réseaux de neurones (Irsoy and Cardie (2014)). On assiste maintenant à une évolution de la fouille d'opinion, qui initialement avait comme objectif l'assignation d'une polarité à un texte, vers des types d'analyse plus complexes, impliquant non seulement la polarité mais aussi les émotions, et la détermination des cibles des émotions ou encore la détermination des aspects sur lesquels sentiments et émotions portent.

Néanmoins, il reste des défis importants à résoudre dans cet axe:

- détecter des émotions spécifiques (par exemple, peur ou rage) et leur intensité. Dans ce cas, il faut aussi savoir faire la distinction entre l'émotion exprimé par celui qui écrit le message et l'émotion suscité par un message;
- identifier l'utilisation de langage figuratif (par exemple sarcasme ou ironie)
- la détection d'attitude ("stance"): déterminer si un orateur est à faveur ou pas d'une certaine proposition (Sobhani et al. (2015))
- l'analyse d'une opinion (générale, plus ou moins partagée), d'un sentiment (personnel, construit avec le temps), d'une émotion (immédiate, pouvant aboutir à des sentiments contradictoires)
- analyse de la dynamique des opinions, sentiments, émotions
- croisement de différentes modalités exprimant des émotions (texte écrit, oral)

L'exemple suivant illustre la complexité de la tâche:
orateur A : les migrants ont le droit de chercher un meilleur endroit pour vivre.
orateur B : c'est pour ça que 25% de la population pénitentiaire est composée d'étrangers...

On peut déduire du texte que B est contre la position soutenue par A. Cependant, c'est une tâche difficile pour les ordinateurs, car il peut nécessiter de raisonnement et/ou de l'intégration de connaissances qui ne se trouvent pas dans le texte

4 Thèmes non encore couverts

Les thèmes suivant sont préssentis pour appartenir à cet axe, mais n'ont pas encore été couverts.

- Lexicométrie
- Annotation de documents
- Veille
- Synthèse et agrégation de documents
- Cartographie d'un domaine (usage des termes, des concepts, évolution temporelle, tendances...), visualisation

- Profilage de textes, attribution d'auteurs, personnalisation linguistique (Personne préssentie : Cyril Labbé)

5 Synthèse à faire à partir des contributions

5.1 Thèmes et périmètre d'étude

1/2 page Présentation de l'axe de réflexion, son périmètre, ses thématiques, ses applications phare

5.2 Etat des lieux

1 à 2 pages (état scientifique)

Pour chaque thème où en est d'un point de vue international Les travaux de référence, logiciels, Expertise disponible en France

5.3 Grands enjeux

1 à 2 pages

Le projet de recherche.

Quels sont le ou les grands enjeux où on doit aller en fonction des forces existantes en France, où il faudrait aller.

5.4 Positionnement

1/2 page NB: par rapport à ce qui existe déjà comme structures.

Interface Autre GDR, autre institut

5.5 Programmatique

1 page Actions à entreprendre

5.6 autre

un fait marquant en 2018 - une médiation scientifique (une réalisation grand public visible) - un lien avec l'industrie

References

Bast, H., Buchhold, B., Haussmann, E., et al. (2016). Semantic search on text and knowledge bases. *Foundations and Trends® in Information Retrieval*, 10(2-3):119–271.

Chen, T. and Van Durme, B. (2017). Discriminative information retrieval for question answering sentence selection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics:*

Volume 2, Short Papers, pages 719–725. Association for Computational Linguistics.

Das, R., Zaheer, M., Reddy, S., and McCallum, A. (2017). Question answering on knowledge bases and text using universal schema and memory networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 358–365. Association for Computational Linguistics.

Ferrero, J. (2017). *Cross Lingual Semantic Textual Similarity Detection : towards Cross-Language Plagiarism Detection*. Theses, LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES.

Ferrero, J., Agnès, F., Besacier, L., and Schwab, D. (2016). A Multilingual, Multi-Style and Multi-Granularity Dataset for Cross-Language Textual Similarity Detection. In *10th edition of the Language Resources and Evaluation Conference*, Portorož, Slovenia.

Francopoulo, G., Mariani, J., and Paroubek, P. (2016). A study of reuse and plagiarism in lrec papers. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Gibney, E. (2006). I'm No Plagiarist, I Moved a Comma. *The Times Higher Education Supplement: THE*. No. 2104.

Grau, B., Ligozat, A.-L., and Gleize, M. (2015). Recherche d'information précise dans des sources d'information structurées et non structurées: défis, approches et hybridation. *Traitement Automatique des Langues*, 56(3).

Guibert, P. and Michaut, C. (2011). Le plagiat étudiant. *Education et sociétés*, 28:214.

Irsoy, O. and Cardie, C. (2014). Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 720–728.

Josephson Institute (2011). WHAT WOULD HONEST ABE LINCOLN SAY? In *Installment 2: Honesty and Integrity - The Ethics of American Youth: 2010, study by Josephson Institute of Ethics' Report Card on American Youth's Values and Actions*.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

McCabe, D. (2010). Students' cheating takes a high-tech turn.

Savenkov, D. and Agichtein, E. (2017). Evinets: Neural networks for combining evidence signals for factoid question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 299–304.

Sobhani, P., Inkpen, D., and Matwin, S. (2015). From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77, Denver, CO. Association for Computational Linguistics.

Wang, W., Yang, N., Wei, F., Chang, B., and Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198. Association for Computational Linguistics.

Yu, M., Yin, W., Hasan, K. S., dos Santos, C., Xiang, B., and Zhou, B. (2017). Improved neural relation detection for knowledge base question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 571–581. Association for Computational Linguistics.